Detection of Conserved RNA Secondary Structures: Hepatitis B as an Example

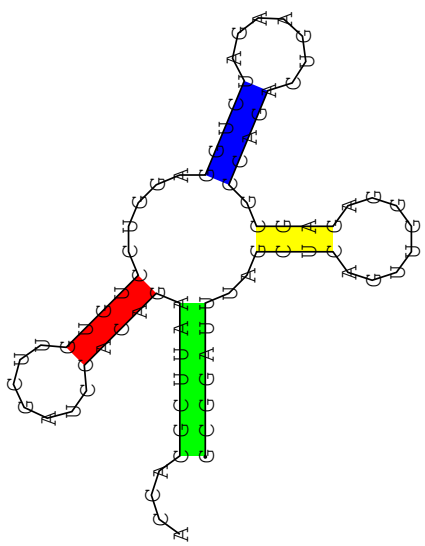
Roman STOCSITS

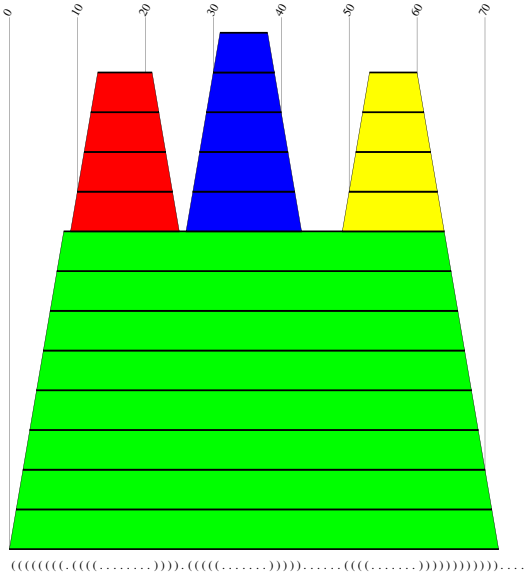
Theoretical Biochemistry Group

University of Vienna

Bled, Jan 2002

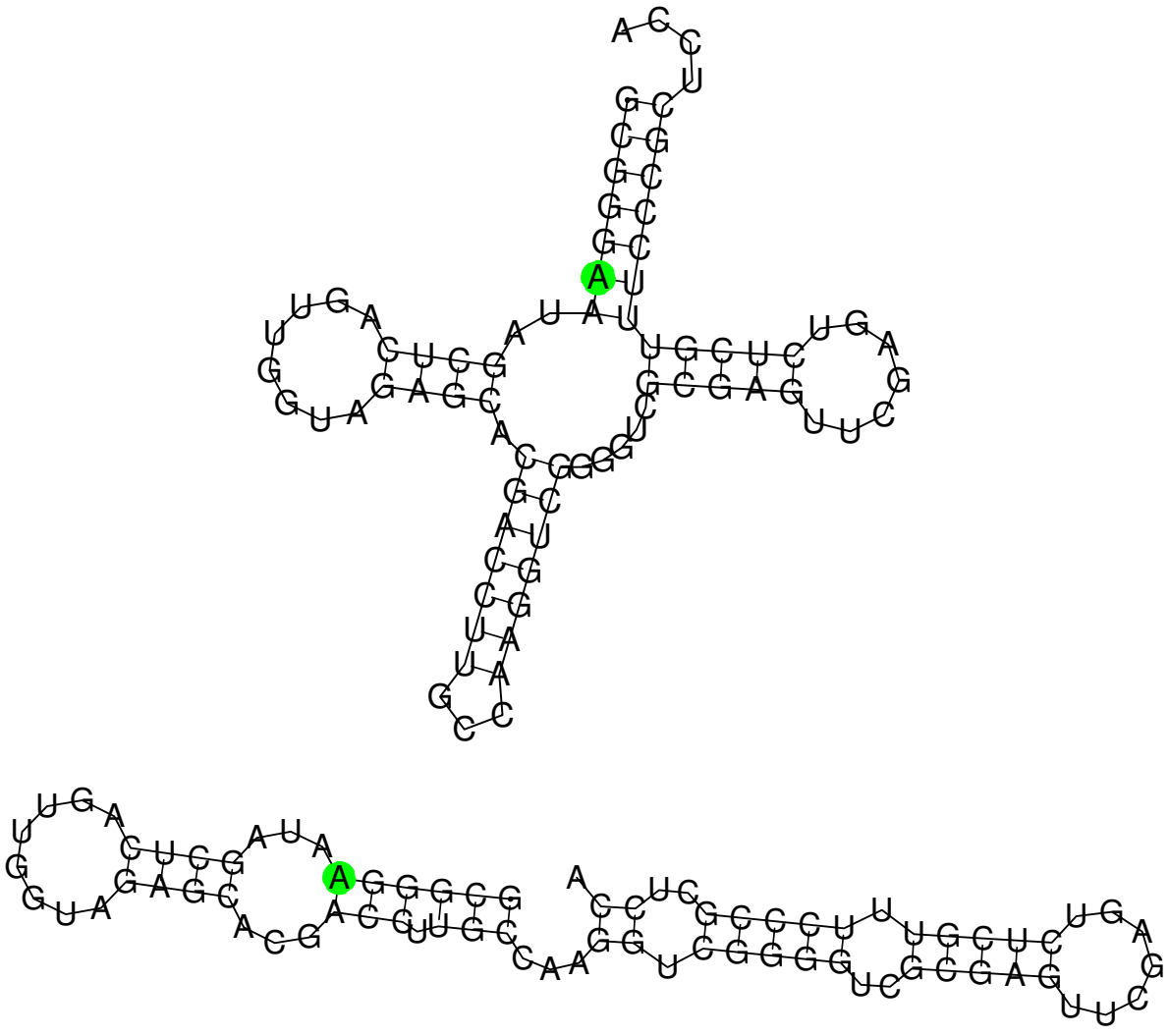
Representation of Secondary Structures





(((((((..((((.........)))).(((((.........)))))(((((.........)))))(((((((....

The Effect of a Single Mutation



```
SARGLSSTVSLGQFEHWSPR
+AR+LS+TVSL+QF+H SPR
NARNLSDTVSLSQFDHPSPR

AGTGCAAGAGGATTAAGTAGTACAGTAAGTTTAGGACAATTTGAACATTGGAGTCCAAGA
   GC  G G   T       AC G      T    CA TT GA CA       CC  G
GACGCCCGCGACCTCTCCGACACCGCTTCCCTCTCCCAGTTCGACCACCCCTCCCCCCGC
```

Example for the problem of higher sequence heterogenity on the level of nucleic acids. Amino acid alignment on top with high degree of similarity.

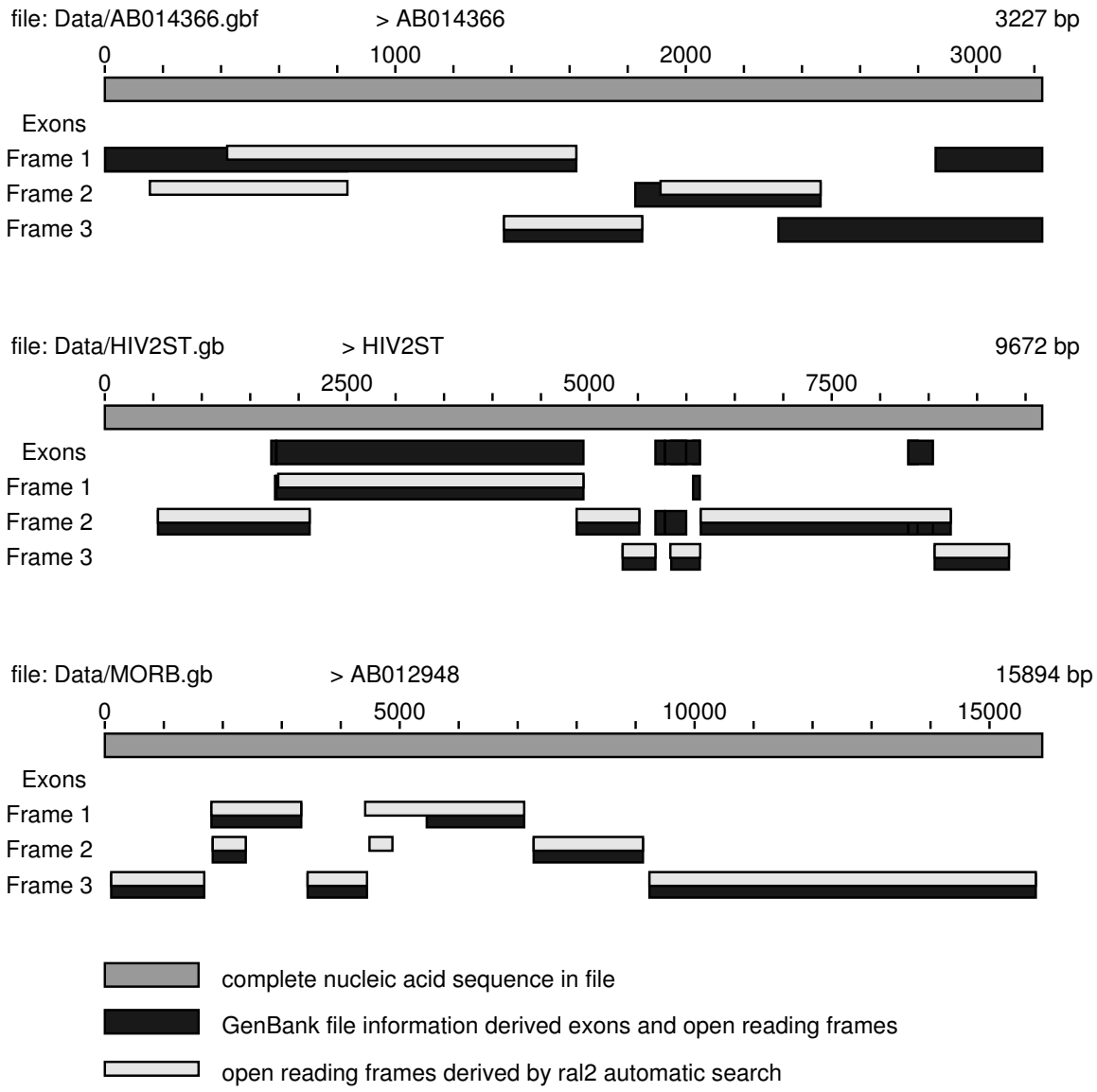
Below the same sequences on level of nucleic acids: much more heterogenous.

Pairwise identity is only 33%.

- Introduction to the problem
- The `RALIGN` project: A short review
- The `RAL2` project: The status quo of development

The following codon tables are available:

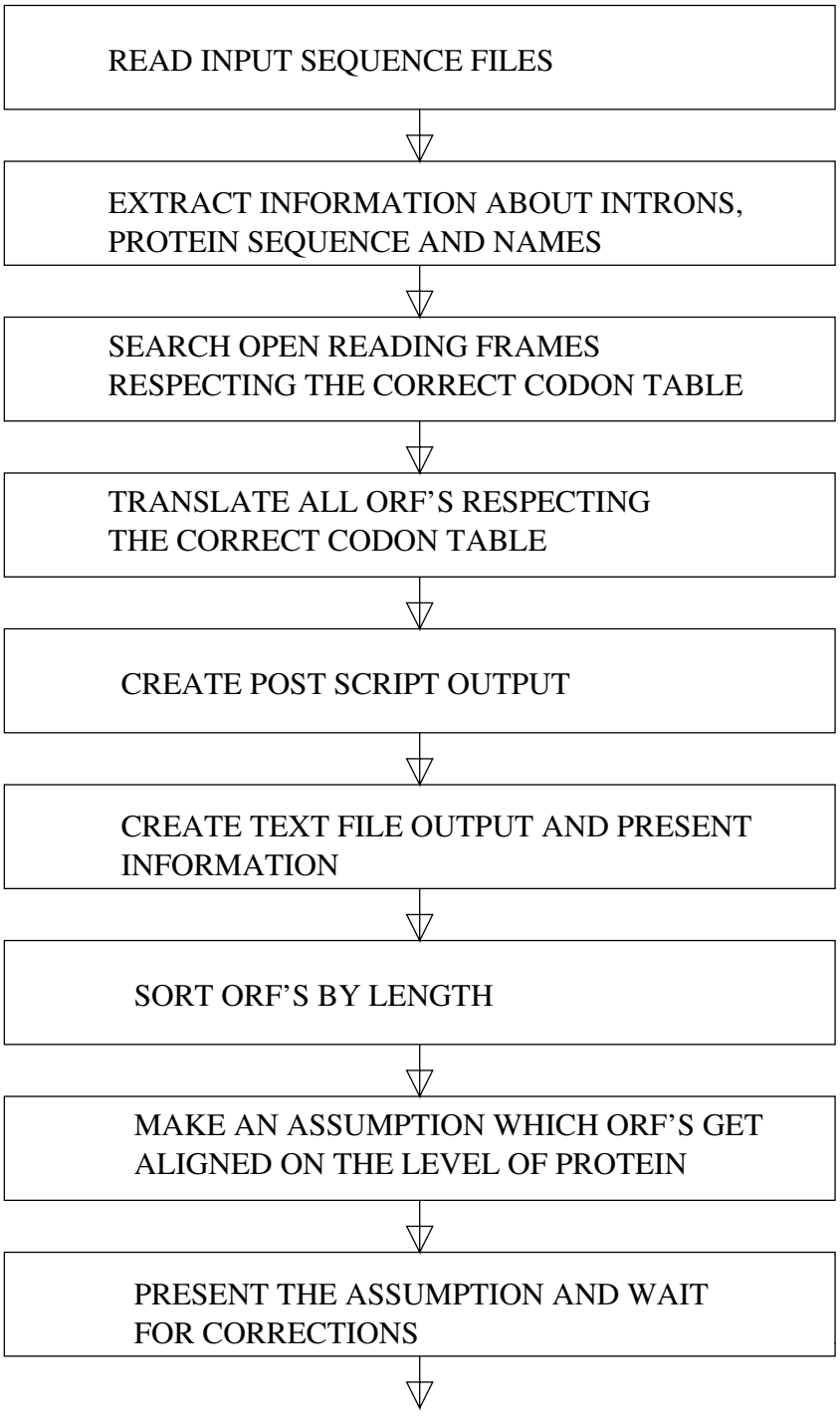
    univ: universal genetic code (default)
    acet: Acetabularia
    ccyl: Candida cylindrica
    tepa: Tetrahymena, Paramecium,
          Oxytrichia, Stylonychia, Glaucoma
    eupl: Euplotes
    mlut: Micrococcus luteus
    mysp: Mycoplasma, Spiroplasma
 mitocan: canonical mitochondrial code
 mitovrt: Vertebrates - mitochondrial code
 mitoart: Arthropods - mitochondrial code
 mitoech: Echinoderms - mitochondrial code
 mitomol: Molluscs - mitochondrial code
 mitoasc: Ascidians - mitochondrial code
 mitonem: Nematodes - mitochondrial code
 mitopla: Plathelminths - mitochondrial code
 mitoyea: Yeasts - mitochondrial code
 mitoeua: Euascomycetes - mitochondrial code
 mitopro: Protozoans - mitochondrial code



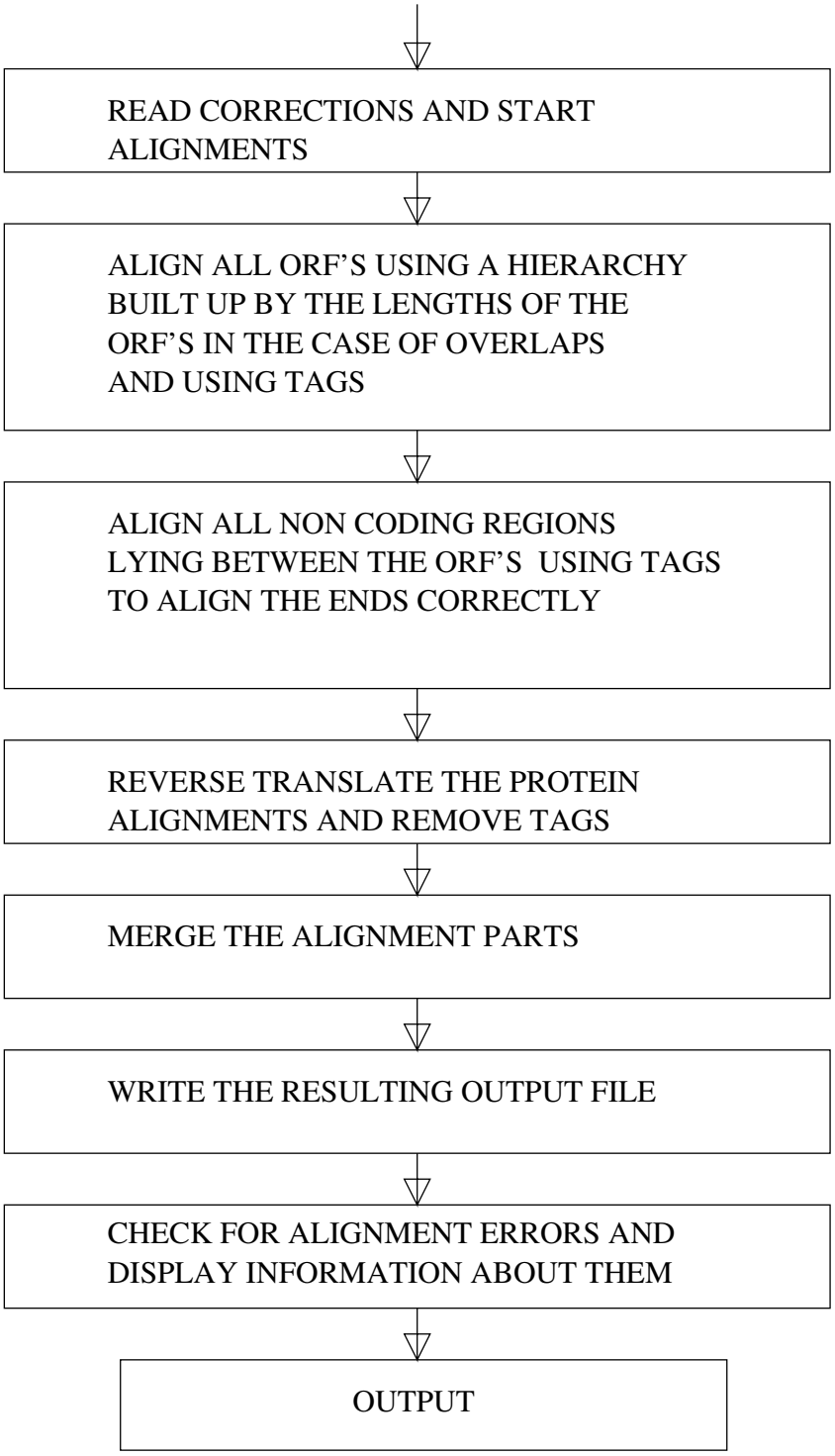
An example for the `PostScript` output of `RAL2`. The figure shows a graphical representation of the found open reading frames of three unrelated sequences. The genomes of hepatitis B virus, HIV1 and the measles virus.

```
File name: Data/AB014366.gbf
Codon Table: univ
Sequence name: > AB014366   [3227bp]
GenBank file information derived exons, if available:
  no exons
GenBank file derived CDS marked open reading frames:
  START - STOP:
  1374 - 1850
  1826 - 2464
  2319 - 3227
  1 - 1623
  2860 - 3227
  1 - 835
Open reading frames derived by ral2 automatic search:
  START - STOP:
  155 - 835
  421 - 1623
  1374 - 1850
  1913 - 2464
```

An example for the text file output of RAL2. Various data about the input sequences are shown: the names of files and sequences, the codon table used, the start and stop codons of found and Genbank derived open reading frames, and, if available, the exon data.



This flow chart shows the first main steps of `RALIGN`.



This flow chart shows the final main steps of `RALIGN`.

OUTPUT:

ral.aln: the resulting alignment file

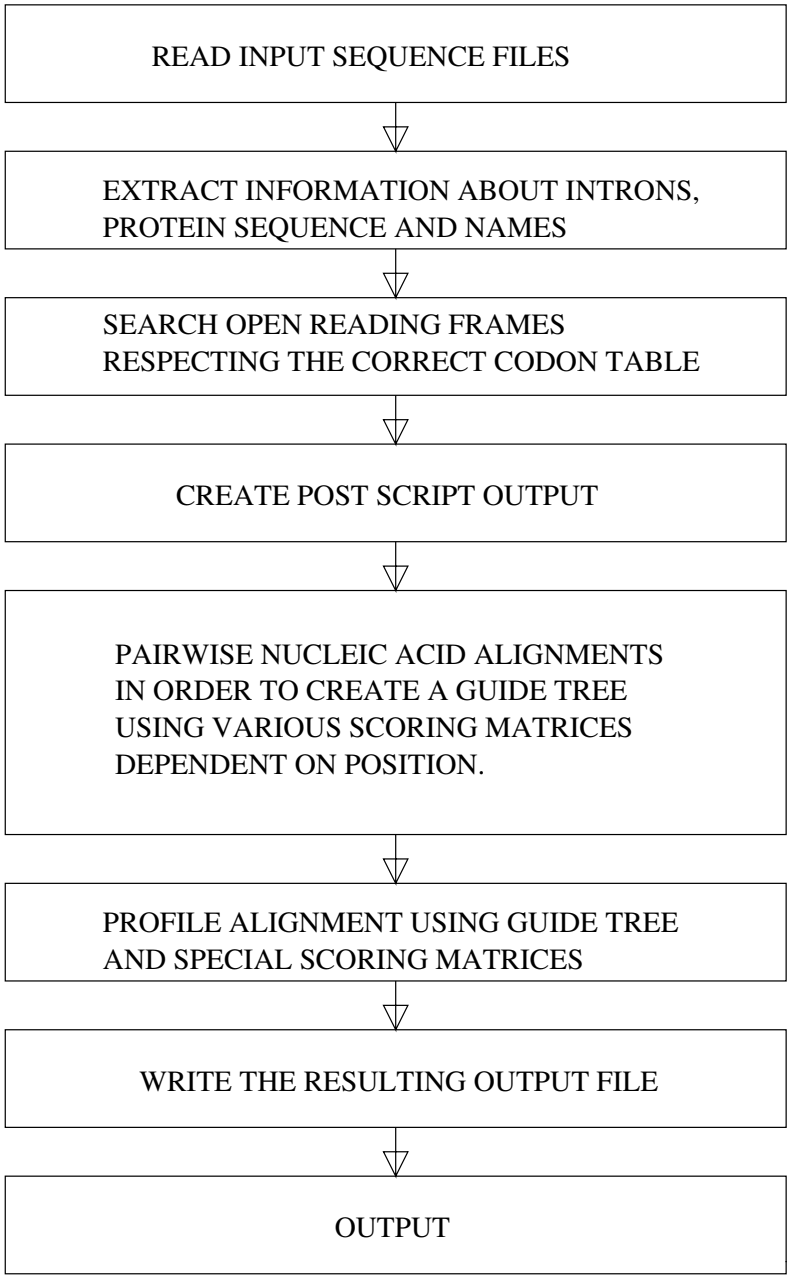
ORF.ps: the Post Script output

ORF.ral: the text output

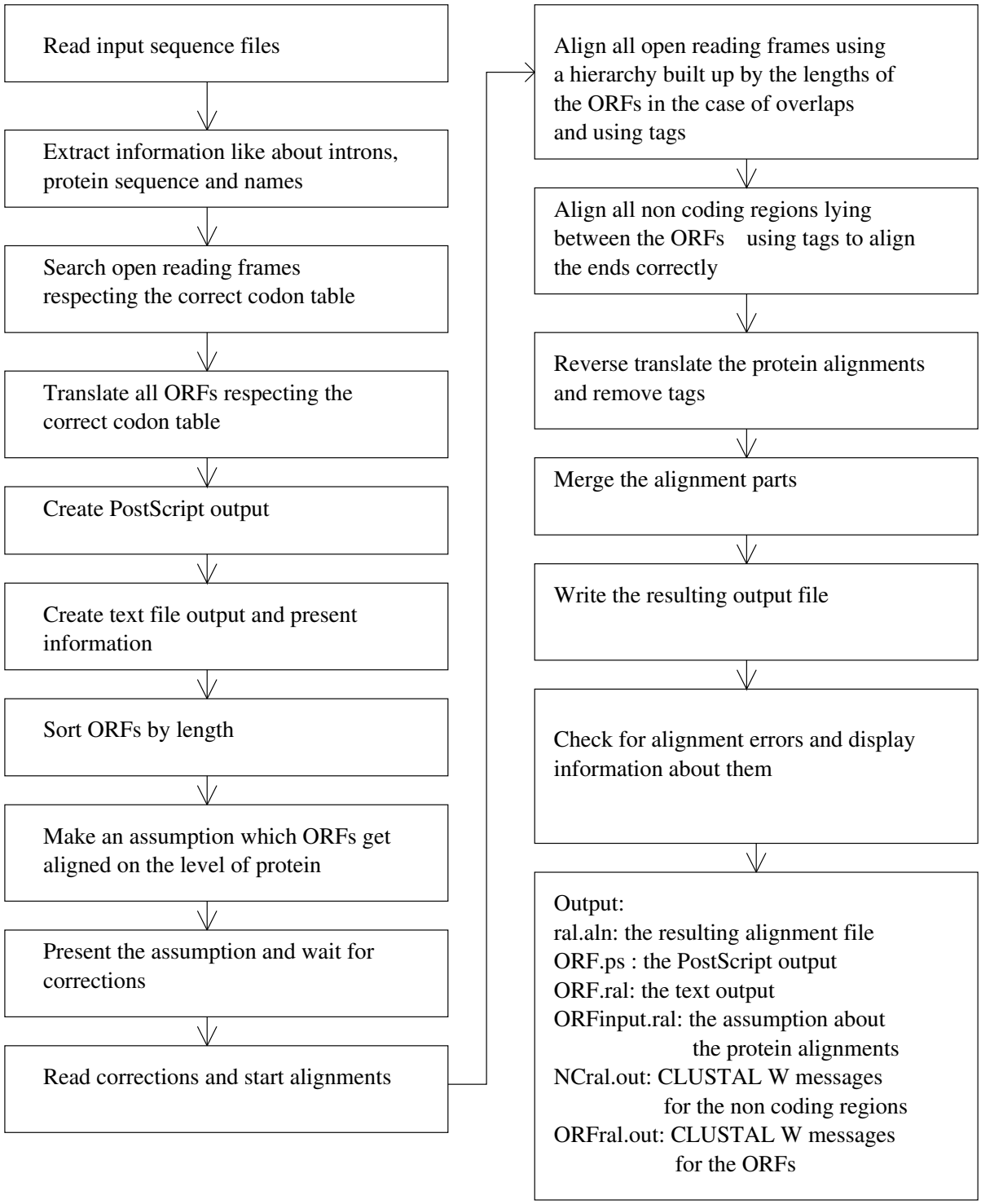
ORFinput.ral: the assumption about the protein alignments

NCral.out: CLUSTALW messages for the non coding regions

ORFral.out: CLUSTALW messages for the ORF's



This flow chart gives a graphical representation of the main steps of `ral2`.



This flow chart shows the main steps of RALIGN.

RALIGN: improved alignments

```
ADI-MAL         UAAACUGCACUAAUGUGAAUGGGACUGCUGU---GAAUGGGACUAAUGCUGGGAGUAAUA
AE-90CF402      UACAUUGUACCAAGG--CU----------------AGUUUUACUAAUGCCA---------
AE-CM240        UAAAUUGUACCAAUG--CU----------------AAUUUGACCAAUGGCAGUAGCAAAA
B-896           UAAAUUGCACUAAUUUGAAUAUCACUA------AGAAUACUACUAAUCCCACUAGUAGCA
B-ACH320A       UAAAUUGCACUGAU---UUUG------------GGAAUGCUACUAAUACCACUAGUAGUA
B-BCSG3         UAAAUUGCACUGAUGAGUUGA------------AGAAUGCUACUAAUACCACUAGUACUA
B-CAM1          UAAUUUGCACUAAUG-UA----------------AAUAAUACUAGGACCAAUAGUAGUG
B-D31           UAAAUUGCACUGAUC-UGAAG------------AAUGCUACUAAUACCAAUAAUAGUA
B-HIV1AD8       UAAAUUGCACUGAUU-UGAGG------------AAUGUUACUAAUAUCAAUAAUAGUA
B-HXB2          UAAAGUGCACUGAU---UUGA------------AGAAUGAUACUAAUACCAAUAGUAGUA
B-JRCSF         UAAAUUGCAAAGAUGUGAA----------------UGCUACUAAUACCACUAGUAGUA
B-LAI           UAAAGUGCACUGAU---UUGG------------GGAAUGCUACUAAUACCAAUAGUAGUA
B-MANC          UAGAUUGCACUGAUUAUGUAG------------GGAAUGCUACUAAUACCACUAGCACUA
B-OYI           UAGAUUGCACUGAUGUUAAUA------------CCACUAGUAGUAGUUUGAGGAAUGCUA
B-SF2           UAAAUUGCACUGAUU-UGGGG------------AAGGCUACUAAUACCAAUAGUAGUA
B-WEAU          UAAAUUGCACUAAUGUGAAUGUGACUAAUUUGAAGAAUGAGACUAAUACCAAUAGUAGUA
B-YU2           UAAAUUGCACUGAU---UUAA------------GGAAUGCUACUAAUACCACUAGUAGUA
B-pNL43         UAAAGUGCACUGAU---UUGA------------AGAAUGAUACUAAUACCAAUAGUAGUA
D-ELI           UAAACUGUAGUGAU--------------------GAAU-UGAGGAACA---AUGGCACUA
D-NDK           UAAACUGCACUGAU--------------------GAAU-UGAGGAACAGCAAGGGCA--A
O-ANT70         UGGAGUGU---------------------------------ACAAACAUAGCUGGAACAA
O-MVP5180       UGAACUGUGUAGAUC---------------U------GCAAACAAAUAAAACAGGCCUAU
SIVCPZGAB       UGCAGUGCAGUAAGG------------------CUAACUUUAGCCAGGCAAAAAACCUAA

ADI-MAL         UAAACUGCACUAAUGUGAAUGGGACUGCUGUGAAUGGGACUAAUGCUGGGAGUAAUAGGA
AE-90CF402      UACAUUGUACCAAG----------------------------GCUAGUUUUACUAAUGCCA
AE-CM240        UAAAUUGUACCAAU----------------------------GCUAAUUUGACCAAUGGCA
B-896           UAAAUUGCACUAAUUUGAAUAUC---------------ACUAAGAAUACUACUAAUCCCA
B-ACH320A       UAAAUUGCACUGAUUUUGGG-----------------------AAUGCUACUAAUACCA
B-BCSG3         UAAAUUGCACUGAU----------------------GAGUUGAAGAAUGCUACUAAUACCA
B-CAM1          UAAUUUGCACUAAUGUAAAU---------------AAUACUAGGACCAAUAGUAGUGAUU
B-D31           UAAAUUGCACUGAUCUGAAG------------------------AAUGCUACUAAUACCA
B-HIV1AD8       UAAAUUGCACU------------------------GAUUUGAGGAAUGUUACUAAUAUCA
B-HXB2          UAAAGUGCACUGAUUUGAAG------------------------AAUGAUACUAAUACCA
B-JRCSF         UAAAUUGCAAAGAUGUG--------------------------AAUGCUACUAAUACCA
B-LAI           UAAAGUGCACUGAUUUGGGG-----------------------AAUGCUACUAAUACCA
B-MANC          UAGAUUGCACUGAUUAUGUAGGG--------------------AAUGCUACUAAUACCA
B-OYI           UAGAUUGCACUGAUGUUAAUACCACUAGUAGU---AGUUUGAGGAAUGCUACUAAUACCA
B-SF2           UAAAUUGCACUGAU-------------------------UUGGGGAAGGCUACUAAUACCA
B-WEAU          UAAAUUGCACUAAUGUGAAUGUGACU---------AAUUUGAAGAAUGAGACUAAUACCA
B-YU2           UAAAUUGCACUGAU-------------------------UUAAGGAAUGCUACUAAUACCA
B-pNL43         UAAAGUGCACUGAUUUGAAG------------------------AAUGAUACUAAUACCA
D-ELI           UAAACUGUAGUGAUGAAUUGAGGAACAAUGGC---------ACUAUGGGGAACAAUGUCA
D-NDK           UAAACUGCACUGAUGAAUUGAGGAAC---------------AGCAAGGGCAAUGGGAAGG
O-ANT70         UGGAGUGU---------------------------------ACAAACAUAGCUGGAACAA
O-MVP5180       UGAACUGUGUAGAUCUGCAA---------------------ACAAAUAAAACAGGCCUAU
SIVCPZGAB       UGCAGUGCAGUAAGGCUAACUUUAGC---------CAGGCAAAAAACCUAACAAACCAGA
```

44% less gaps

RALIGN: improved alignments

```
ADI-MAL         GGGAGUAAUAGGACUA-------AUGCAG--AAUUGAAA--AUGGAAAUUGGAGAAGUGAAAAAC
AE-90CF402      A---------CCAGUG-----------ACAGAAUA---AAAAUG---GAAGAUGCAGUAAGAAAC
AE-CM240        AGUAGCAAAACCAAUGU-----CUCUAACAUAAUAGGAAAUAUA---ACAGAUGAAGUAAGAAAC
B-896           ACUAGUAGCA---------------GCUGGGGAAUGA-----UGGAGAAAGGAGAAAUAAAAAAU
B-ACH320A       ACUAGUAGUA---------------GCGGGGUUAUAAUA----G-AGAAAGGAGAAAUAAAAAAC
B-BCSG3         ACUAGUACUAAUACCCCUAGUGGUAGCUGGAAAAAGAU-----GGAAAGAGGAGAAAUAAAGAAC
B-CAM1          AAUAGUAGUGAUU-----GGGACAGGAGGGAAGGAGAAAAGAUG---AAAGGAGAAAUAAAAAAC
B-D31           AAUAAUAGUAGU-----------UGGACGAUGACAGGAGAAAUG---AAAGGAGAAAUAAAAAAC
B-HIV1AD8       AAUAAUAGUAGU-------------GAGGGAA----------UG---AGAGGAGAAAUAAAAAAC
B-HXB2          AAUAGUAGUA---------------GCGGGAGAAUGAUA--AUGGAGAAAGGAGAGAUAAAAAAC
B-JRCSF         ACUAGUAGUA---------------GUGAGGGAAUGA-----UGGAGAGAGGAGAAAUAAAAAAC
B-LAI           AAUAGUAGUAAUACCAAUAGUAGUAGCGGGGAAAUGAUG--AUGGAGAAAGGAGAGAUAAAAAAC
B-MANC          ACUAGCACUAAUAAUACCGCUAGUGGAAGUUGG--GGAGCGAUG---AGAGGGGAAAUAAAAAAC
B-OYI           AGGAAUGCUACUAAUACCACAAGUAGUAGUUGG--GAAACGAUGGAGAAAGGAGAAUUAAAAAAC
B-SF2           AAUAGUAGUAAUU-----GGAAA--GAAGAAA----------UA---AAAGGAGAAAUAAAAAAC
B-WEAU          AAUAGUAGUAGU------------GGAGGGGAAAAGA-----UGGAGGAGGGAGAAAUGAAAAAC
B-YU2           ACUAGUAGUA---------------GCUGGGAAACGAU-----GGAGAAAGGAGAAAUAAAAAAC
B-pNL43         AAUAGUAGUA---------------GCGGGAGAAUGAUA--AUGGAGAAAGGAGAGAUAAAAAAC
D-ELI           -AUGGCACUA-U-------------GGGGAACAAUGUCACUACAGAGGAGAAAGGAAUGAAAAAC
D-NDK           AAGGGCA--A-U-------------GGG----AAGGUA---GAAGAGGAGGAAAAAAGGAAAAAC
O-ANT70         GCUGGAACAA--------------------CAAAUGA------------AAACCUUAUGAAGAAG
O-MVP5180       ACAGGCCUAU--------------------UAAAUGAGA------CAAUAAAUGAGAUGAGAAAU
SIVCPZGAB       AAAAACCUAA---------------CAAACCAGACAUCUUCUC-CGCCUCUCGAAAUGAAAAAC
```

```
ADI-MAL         ACUAAUGCAGAAUUGAAAAUGGAA---------------------AUUGGAGAAGUGAAAAAC
AE-90CF402      ACCAGUGACAGA-------------------------AUAAAAAUGGAAGAUGCAGUAAGAAAC
AE-CM240        AGUAGCAAAACCAAUGUCUCU------AACAUAAUAGGAAAUAUAACAGAUGAAGUAAGAAAC
B-896           ACUAGUAGCAGC------------UGGGGA------AUGAUGGAGAAAGGAGAAAUAAAAAAU
B-ACH320A       ACUAGUAGU---------------AGCGGG---GUUAUAAUAGAGAAAGGAGAAAUAAAAAAC
B-BCSG3         ACUAGUACUAAU---ACCCCUAGUGGUAGCUGGAAAAAGAUGGAAAGAGGAGAAAUAAAGAAC
B-CAM1          UGGGACAGGAGGGAAGGAGAA---------------AAG---AUGAAAGGAGAAAUAAAAAAC
B-D31           AAUAAUAGUAGUUGGACGAUGACAGGA------------GAAAUGAAAGGAGAAAUAAAAAAC
B-HIV1AD8       AAUAAUAGUAGUGAGGGA-----------------------AUGAGAGGAGAAAUAAAAAAC
B-HXB2          AAUAGUAGU---------------AGCGGGAGAAUGAUAAUGGAGAAAGGAGAGAUAAAAAAC
B-JRCSF         ACUAGUAGUAGU------------GAGGGA------AUGAUGGAGAGAGGAGAAAUAAAAAAC
B-LAI           AAUAGUAGUAAUACCAAUAGUAGUAGCGGGGAAAUGAUGAUGGAGAAAGGAGAGAUAAAAAAC
B-MANC          ACUAGCACUAAUAAUACCGCUAGUGGAAGUUGG---GGAGCGAUGAGAGGGGAAAUAAAAAAC
B-OYI           ACAAGUAGU-----------------AGUUGGGAAACGAUGGAGAAAGGAGAAUUAAAAAAC
B-SF2           AAUAGUAGU-----------------AAUUGGAAAGAAGAAAUAAAAGGAGAAAUAAAAAAC
B-WEAU          AAUAGUAGUAGUGGAGGGGAA---------------AAGAUGGAGGAGGGAGAAAUGAAAAAC
B-YU2           ACUAGUAGU-----------------AGCUGGGAAACGAUGGAGAAAGGAGAAAUAAAAAAC
B-pNL43         AAUAGUAGU---------------AGCGGGAGAAUGAUAAUGGAGAAAGGAGAGAUAAAAAAC
D-ELI           ACUACAGAGGAGAAAGGAAUG-----------------------------------AAAAAC
D-NDK           GUAGAAGAGGAGGAAAAAAGG-----------------------------------AAAAAC
O-ANT70         ACAAAUGAAAACCUU-----------------------------------------AUGAAGAAG
O-MVP5180       UUAAAUGAGACAAUAAAU--------------------------------GAGAUGAGAAAU
SIVCPZGAB       ACAUCUUCUCCG--------------------------------CCUCUCGAAAUGAAAAAC
```

50% less gaps

RALIGN: improved alignments

```
ADI-MAL         CAAAU----AG-----------------AUGAUAGUGAUAAUAG--------AUAG---UACUAAUUAUAGGUUAAUAAAUUG
AE-90CF402      CAAAUUAGUAGUAGUGUACAAAAUAAUAAUAACAGUAAUACUAGUGGACAAAAUAAUAGUCAUAAGUUUAGAUUAAUACAUUG
AE-CM240        CAAAUU-GAAG-----------------AUAAGAAGACUAGUAGUG-----------------AGUAUAGGUUAAUAAAUUG
B-896           CCAAU----AG-----------------AAAAUACUAAUAAUA------------------CUAAGUAUAGGUUAAUAAGUUG
B-ACH320A       CCAAU----AG-----------------AUAAUAAUAAUACUAAUA------CCAGCUAUACCAGCUAUAGGUUGAUAAGUUG
B-BCSG3         CCAAU----AG-----------------AUAAUGAUAAGAAUAG--------------UACCAAAUAUAGGUUGAUAAGUUG
B-CAM1          CCAAU----AG-----------------AUAAGGCUAAUACAAGU---------------------UAUACAUUGAUACAUUG
B-D31           CCAAU----AG-----------------AUAAUGACAAUACUAG---------------------CUAUAGGUUGAUAAGUUG
B-HIV1AD8       CCAAU----AG-----------------AUAAUGAUAAUACUAG---------------------CUAUAGGUUGAUAAAUUG
B-HXB2          CCAAU----AG-----------------AUAAUGAUAC----------------------UACCAGCUAUAAGUUGACAAGUUG
B-JRCSF         CCAAU----AG-----------------AUAAUAAGAAUAAUA------------------CCAAAUAUAGGUUAAUAAGUUG
B-LAI           CCAAU----AG-----------------AUAAUGAUAC----------------------UACCAGCUAUACGUUGACAAGUUG
B-MANC          CCAAU----AG-----------------AAAAGAAGAAUACUAG---------------------CUUUAGAUUGAUAAGUUG
B-OYI           CCAAU----AG-----------------AUAAGAAUGAUACUAA---------------------AUUUAGGUUAAUACAUUG
B-SF2           CCAAU----AG-----------------AUAAUGCUAGUACUACUA------CCAACUAUACCAACUAUAGGUUGAUACAUUG
B-WEAU          CCAAU----AG-----------------AUCAUGAUAAUACAAG---------------------CUAUACGUUGAUAAAUUG
B-YU2           CCAAU----AG-----------------AUAAUG---------------------------CUAGCUAUAGGUUGAUAAGUUG
B-pNL43         CCAAU----AG-----------------AUAAU----------------------ACCAGCUAUAGGUUGAUAAGUUG
D-ELI           CCAAU----AG-----------------ACAAUGAUAGUAGUACCA------AUAG---UACCAAUUAUAGGUUAAUAAAUUG
D-NDK           CCAAU----AG-----------------ACAAUAAUAAUAGGACCA------AUAG---UACUAAUUAUAGGUUAAUAAAUUG
O-ANT70         GAACUG--AAUGAG----ACAAGCAGCACAAAUAAGACAAACAGC---------------AAAAUGUAUACAUUAACUAAUUG
O-MVP5180       AAGGUU--AAUGA-------------CUCAAAUGCAGUAAAUGGA---------------ACAACAUAUAUGUUAACUAAUUG
SIVCPZGAB       AACCU----AG-----------------GGAAUGAGAACAACAC---------------------AUAUAGGAUAAUUAAUUG


ADI-MAL         CAAAUAGAUGAU-----------------------------------------AGUGAUAAUAGUAGUUAUAGGCUAAUAAAUUG
AE-90CF402      CAAAUUAGUAGUAGUGUACAAAAUAAUAAUAACAGUAAUACUAGUGGACAAAAUAAUAGUCAUAAGUUUAGAUUAAUACAUUG
AE-CM240        CAAAUUGAA---------------------GAU----------------AAGAAGACUAGUAGUGAGUAUAGGUUAAUAAAUUG
B-896           CCAAUAGAAAAUACUAAU-------------------------------------AAUACUAAGUAUAGGUUAAUAAGUUG
B-ACH320A       CCAAUAGAUAAUAAUAAUACUAAU--------------------------ACCAGCUAUACCAGCUAUAGGUUGAUAAGUUG
B-BCSG3         CCAAUAGAUAAUGAUAAG-----------------------------------AAUAGUACCAAAUAUAGGUUGAUAAGUUG
B-CAM1          CCAAUAGAUAAGGCU--------------------------------------AAUACAAGUUAUACAUUGAUACAUUG
B-D31           CCAAUAGAUAAUGAC--------------------------------------AAUACUAGCUAUAGGUUGAUAAGUUG
B-HIV1AD8       CCAAUAGAUAAUGAU--------------------------------------AAUACUAGCUAUAGGUUGAUAAAUUG
B-HXB2          CCAAUAGAUAAUGAU--------------------------------------ACUACCAGCUAUAAGUUGACAAGUUG
B-JRCSF         CCAAUAGAUAAUAAGAAU-----------------------------------AAUACCAAAUAUAGGUUAAUAAGUUG
B-LAI           CCAAUAGAUAAUGAU--------------------------------------ACUACCAGCUAUACGUUGACAAGUUG
B-MANC          CCAAUAGAAAAGAAG--------------------------------------AAUACUAGCUUUAGAUUGAUAAGUUG
B-OYI           CCAAUAGAUAAGAAU--------------------------------------GAUACUAAAUUUAGGUUAAUACAUUG
B-SF2           CCAAUAGAUAAUGCUAGUACUACU-------------------------ACCAACUAUACCAACUAUAGGUUGAUACAUUG
B-WEAU          CCAAUAGAUCAUGAU--------------------------------------AAUACAAGCUAUACGUUGAUAAAUUG
B-YU2           CCAAUAGAUAAU-----------------------------------------GCUAGCUAUAGGUUGAUAAGUUG
B-pNL43         CCAAUAGAUAAU-----------------------------------------ACCAGCUAUAGGUUGAUAAGUUG
D-ELI           CCAAUAGACAAUGAUAGUAGU----------------------------ACCAAUAGUACCAAUUAUAGGUUAAUAAAUUG
D-NDK           CCAAUAGACAAUAAUAAUAGG----------------------------ACCAAUAGUACUAAUUAUAGGUUAAUAAAUUG
O-ANT70         GAACUGAAUGAGACAAGCAGCACAAAUAAGACAAACAGC---------------AAAAUGUAUACAUUAACUAAUUG
O-MVP5180       AAGGUUAAUGACUCAAAUGCAGUAAAUGGAACAACA-----------------------UAUAUGUUAACUAAUUG
SIVCPZGAB       AACCUAGGGAAUGAG--------------------------------------AACAACACAUAUAGGAUAAUUAAUUG
```

66% less gaps

RALIGN: improved alignments

```
ADI-MAL       CUCUAUACAACAGGGAUAGU--------A----GGAG--AUAU-AAGAAGAGCAUAUUGUACU
AE-90CF402    UUCCAUACAACAGGAAACAU-------AAAU--GGUG--AUAU-AAGAAAAGCAUAUUGUGAA
AE-CM240      UUCUAUAGAACAGGAGAUAU-------AAUA--GGAA--AUAU-AAGAAAAGCAUAUUGUGAG
B-896         UUUUAUGCAAGAAGAAACAU-------AAUA--GGAG--AUAU-AAGACAAGCACAUUGUAAC
B-ACH320A     UUUUAUGCAACAGGACAAAU-------AAUA--GGAG--AUAU-AAGACAAGCACAUUGUAAC
B-BCSG3       UAUUAUACAACAGGAGAAAU-------AGUA--GGAG--AUAU-AAGACAAGCACAUUGUAAC
B-CAM1        GUUUAUGCAACAGACAGAAU-------AAUA--GGAG--AUAU-AAGACAAGCACAUUGUAAC
B-D31         UUUUAUACAAAAGGAAAAAU-------AAUA--GGAG--AUAU-AAGACAAGCACAUUGUAAC
B-HIV1AD8     UUUUAUACAACAGGAGACAU-------AAUA--GGAG--AUAU-AAGACAAGCACAUUGCAAC
B-HXB2        UUUGUUACAAUAGGAAAAAU-------AG-----GAA--AUAU-GAGACAAGCACAUUGUAAC
B-JRCSF       UUUUAUACAACAGGAGAAAU-------AAUA--GGAG--AUAU-AAGACAAGCACAUUGUAAC
B-LAI         UUUGUUACAAUAGGAAAAAU-------AG-----GAA--AUAU-GAGACAAGCACAUUGUAAC
B-MANC        UUUCAUGUAACAAGAGCCGU-------AACA--GGAG--AUAU-AAGACAAGCACAUUGUAAC
B-OYI         UUUCAUACAACAAAACAAAU-------AAUA--GGAG--AUAU-AAGACAAGCACAUUGUAAC
B-SF2         UUUCAUACAACAGGAAGAAU-------AAUA--GGAG--AUAU-AAGAAAAGCACAUUGUAAC
B-WEAU        CUUUAUACAACAGGAGAAAU-------AAUA--GGAG--AUAU-AAGACGAGCACAUUGUAAC
B-YU2         UUGUAUACAACAGGAGAAAU-------AAUA--GGAG--AUAU-AAGACAAGCACAUUGUAAC
B-pNL43       UUUGUUACAAUAGGAAAAAU-------AG-----GAA--AUAU-GAGACAAGCACAUUGUAAC
D-ELI         CUCUAUACUACAAGAUCAA----GAUCAAUA----------AU-AGGACAAGCACAUUGUAAU
D-NDK         CUCUAUACAAUAAACAGGAAAAAAGAAGAAAACAGGAU--ACAU-AGGACAAGCACAUUGUAAA
O-ANT70       UACAGCAUGGGAAU----AGGGGGAACAGCAGGAAAC--AGCUCAAGGGCAGCUUAUUGCAAG
O-MVP5180     CGCAGUAUGACACUUAAAAGAAGUAACAAUACAUCACCAAGAUCAAGGGUAGCUUAUUGUACA
SIVCPZGAB     UUUUAUAAUAUAGAAAAUGU-------AGUA--GGAG--AUAC-CAGAUCUGCCUACUGUAAG
```

```
ADI-MAL         CUCUAUACAACAGGGAUAGUAGGA---------------GAUAUAAGAAGAGCAUAUUGUACU
AE-90CF402      UUCCAUACAACAGGAAACAUAAAUGGU------------GAUAUAAGAAAAGCAUAUUGUGAA
AE-CM240        UUCUAUAGAACAGGAGAUAUAAUAGGA------------AAUAUAAGAAAAGCAUAUUGUGAG
B-896           UUUUAUGCAAGAAGAAACAUAAUAGGA------------GAUAUAAGACAAGCACAUUGUAAC
B-ACH320A       UUUUAUGCAACAGGACAAAUAAUAGGA------------GAUAUAAGACAAGCACAUUGUAAC
B-BCSG3         UAUUAUACAACAGGAGAAAUAGUAGGA------------GAUAUAAGACAAGCACAUUGUAAC
B-CAM1          GUUUAUGCAACAGACAGAAUAAUAGGA------------GAUAUAAGACAAGCACAUUGUAAC
B-D31           UUUUAUACAAAAGGAAAAAUAAUAGGA------------GAUAUAAGACAAGCACAUUGUAAC
B-HIV1AD8       UUUUAUACAACAGGAGACAUAAUAGGA------------GAUAUAAGACAAGCACAUUGCAAC
B-HXB2          UUUGUUACAAUAGGAAAA---AUAGGA------------AAUAUGAGACAAGCACAUUGUAAC
B-JRCSF         UUUUAUACAACAGGAGAAAUAAUAGGA------------GAUAUAAGACAAGCACAUUGUAAC
B-LAI           UUUGUUACAAUAGGAAAA---AUAGGA------------AAUAUGAGACAAGCACAUUGUAAC
B-MANC          UUUCAUGUAACAAGAGCCGUAACAGGA------------GAUAUAAGACAAGCACAUUGUAAC
B-OYI           UUUCAUACAACAAAACAAAUAAUAGGA------------GAUAUAAGACAAGCACAUUGUAAC
B-SF2           UUUCAUACAACAGGAAGAAUAAUAGGA------------GAUAUAAGAAAAGCACAUUGUAAC
B-WEAU          CUUUAUACAACAGGAGAAAUAAUAGGA------------GAUAUAAGACGAGCACAUUGUAAC
B-YU2           UUGUAUACAACAGGAGAAAUAAUAGGA------------GAUAUAAGACAAGCACAUUGUAAC
B-pNL43         UUUGUUACAAUAGGAAAA---AUAGGA------------AAUAUGAGACAAGCACAUUGUAAC
D-ELI           CUCUAUACUACAAGAUCAAGAUCA----------------AUAAUAGGACAAGCACAUUGUAAU
D-NDK           CUCUAUACAAUAACAGGAAAAAAGAAGAAAACAGGA---UACAUAGGACAAGCACAUUGUAAA
O-ANT70         UACAGCAUGGGAAUAGGGGGAACAGCAGGAAACAGC------UCAAGGGCAGCUUAUUGCAAG
O-MVP5180       CGCAGUAUGACACUUAAAAGAAGUAACAAUACAUCACCAAGAUCAAGGGUAGCUUAUUGUACA
SIVCPZGAB       UUUUAUAAUAUAGAAAAUGUAGUAGGA------------GAUACCAGAUCUGCCUACUGUAAG
```

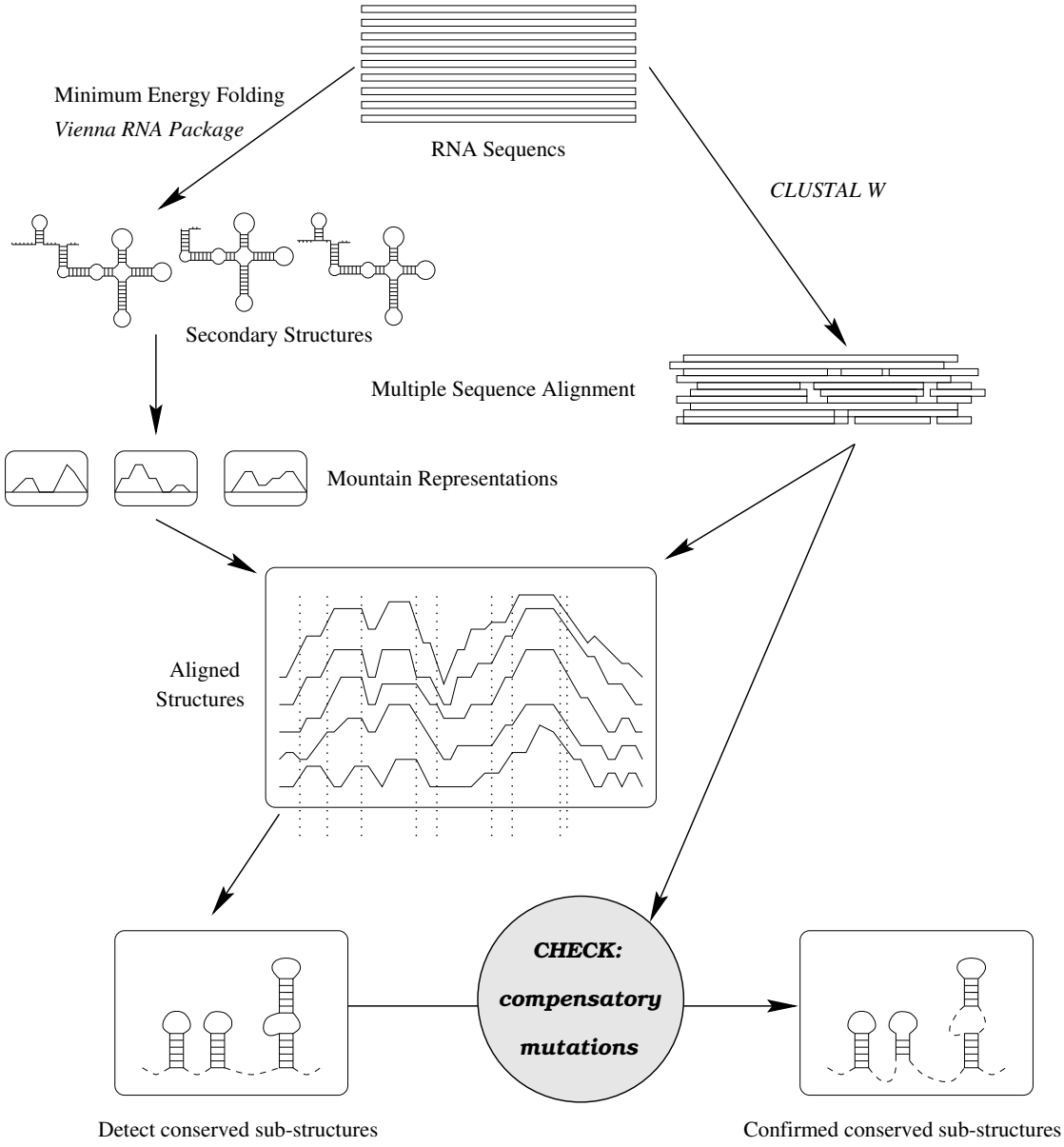
70% less gaps

RALIGN: improved alignments

```
ADI-MAL         AU-AGUACAUGGCAGAAUAAUGGUGC----AAGA------CU-AA--GU----AAUAGCACAGAGUC---AACUGGUAGUAUCACACUCCCAUG
AE-90CF402      AU-AGUACUUGGAUA---------------AAUGGAACCAUGCAGGAGGUU--AAUGGCACAAACUC---A---GGCAAUAUCACACUUCCAUG
AE-CM240        AU-AAUACUUGCCUAG------------GAAAUGAAACCAUGGCGGGGUGU--AAUGACAC-----------------UAUCACACUUCCAUG
B-896           AU-AGUACUUGGAAU-------G-------U-UA------CUGGAGGGACA--AAUGGCACUGAAGG---AAAUGACAUAAUCACACUCCAAUG
B-ACH320A       AU-AGUACUUGG------AAUGAUACUGGGAAUGUUA---CUGAAAGGUCA--AAUAACAAUGA------AAAU------AUCACACUCCCAUG
B-BCSG3         AU-AGUACUUGGGCUGGGAAUAAUACUUGGAAUAGUAGUGCUGAAAGGUCA--GAUGACACUGGAGG---AAAU------AUCACACUCCCAUG
B-CAM1          AU-ACUACUUGGCUGUUUAAUGGUACUUGGAAUGAUA---CUGAAGGGUUA--AAUAACACUGAAAG---AAAU------AUUACACUUCCAUG
B-D31           AU-AGUACUUGGAAU-------------------GAUA---CUAAAGAGUCA--AAUAACACAAAU---------GGAACUAUCACACUCCCAUG
B-HIV1AD8       AU-AGUACUUGGAAUUUUAAUGGUACUUGGAAUUUAA---CACAAUCG-----AAUGGUACUGAAGG---AAAUGACACUAUCACACUCCCAUG
B-HXB2          AU-AGUACUUGGUUU---AAUAGUACUUGGAGUA------CUGAAGGGUCA--AAUAACACUGAAGG---AAGUGACACAAUCACCCUCCCAUG
B-JRCSF         AU-AGUACUUGGAAU-------G-------A-UA------CUGAAAAGUCA--AGUGGCACUGAAGG---AAAUGACACCAUCAUACUCCCAUG
B-LAI           AU-AGUACUUGGUUU---AAUAGUACUUGGAGUA------CUGAAGGGUCA--AAUAACACUGAAGG---AAGUGACACAAUCACACUCCCAUG
B-MANC          AU-AGUACUUGGAAUACUGGG---------AAUGAUA---CUAGAGAGUCA--AAUGACACAAAUAA---UACUGGAAAUAUCACACUCCCAUG
B-OYI           AU-AGUACUUGGAAU-------------------GAUA---CUACAAGGGCA--AAUAGCACUGAA---------GUAACUAUCACACUCCCAUG
B-SF2           AU-AAUACAUGGAGGUUAAAU--------------------CACACU-G-----AA-GGAACUAAAGG---AAAUGACACAAUCAUACUCCCAUG
B-WEAU          AU-AGUACUUGGCAUGCUAAUGGUACUUGGAAGAAUA---CUGAAGGGGCA--GAUAACAAU------------------AUCACACUCCCAUG
B-YU2           -------CUUGG------AAUGAUACUAGAAA---------------GUUA--AAUAACACUGGAAG---AAAU------AUCACACUCCCAUG
B-pNL43         AU-AGUACUUGGUUU---AAUAGUACUUGGAGUA------CUGAAGGGUCA--AAUAACACUGAAGG---AAGUGACACAAUCACACUCCCAUG
D-ELI           AU-AGUACAUGGAAUAU---UAGUGCAUGGAAUAAUAU---UACAGAGUCA--AAUAAUAGCACAAA---CAC---AAACAUCACACUCCAAUG
D-NDK           AU-AGUACAUGGAAU----------CA--GACUAAUAG---UACAGGGUUC--AAUAAUGGCACAG---------------UCACACUCCCAUG
O-ANT70         AUUA-UACCUUU-UCA----------UGUAACGGAACCACCUGUAGUGUUAGUAAUGUUAGUCAAGG------UAACAAUGGCACUCUACCUUG
O-MVP5180       ACUA-UACUUUUAUCAA--------CUGUACAAAGUCCGGAUGCCAGGAGAUCAAAGGGAGCAAUGAGACCAAUAAAAAUGGUACUAUACCUUG
SIVCPZGAB       CUGACAACAUUA-------------------------------------CA--AAUGGCAUU-----------------AUAAUACUGCCAUG
```

```
ADI-MAL           AUAGUACAUGGCAGAAUAAUGGU---GCAAGACUAAGUAAUAGCACAGAGUCAACUGGU-----------------------AGUAUCACACUCCCAUG
AE-90CF402        AUAGUACUUGGAUA-----------AAUGGAACCAUGCAGGAGGUUAAUGGCACAAACUCA-----------------GGCAAUAUCACACUUCCAUG
AE-CM240          AUAAUACUUGCCUAGGA---------AAUGAAACCAUGGCGGGGUGUAAUGACACU-----------------------AUCACACUUCCAUG
B-896             AUAGUACUUGGAAU-----------------GUUACUGGAGGGACAAAUGGCACUGAAGGA---------------AAUGACAUAAUCACACUCCAAUG
B-ACH320A         AUAGUACUUGGAAUGAUACUGGG------AAUGUUACUGAAAGGUCAAAUAACAAUGAA----------------------AAUAUCACACUCCCAUG
B-BCSG3           AUAGUACUUGGGCUGGG-----------------AAUAAUACUUGGAAUAGUAGUGCUGAAAGGUCAGAUGACACUGGAGGAAAUAUCACACUCCCAUG
B-CAM1            AUACUACUUGGCUG-----------------UUUAAUGGUACUUGGAAUGAUACUGAAGGG---UUAAAUAACACUGAAAGAAAUAUUACACUUCCAUG
B-D31             AUAGUACUUGGAAU-----------------GAUACUAAAGAGUCAAAUAACACAAAU----------------------GGAACUAUCACACUCCCAUG
B-HIV1AD8         AUAGUACUUGGAAU-----------------UUUAAUGGUACUUGGAAUUUAACACAAUCGAAUGGUACUGAAGGAAAUGACACUAUCACACUCCCAUG
B-HXB2            AUAGUACUUGGUUUAAUAGUACU------UGGAGUACUGAAGGGUCAAAUAACACUGAAGGA---------------AGUGACACAAUCACCCUCCCAUG
B-JRCSF           AUAGUACUUGGAAU-----------------GAUACUGAAAAGUCAAGUGGCACUGAAGGA---------------AAUGACACCAUCAUACUCCCAUG
B-LAI             AUAGUACUUGGUUUAAUAGUACU------UGGAGUACUGAAGGGUCAAAUAACACUGAAGGA---------------AGUGACACAAUCACACUCCCAUG
B-MANC            AUAGUACUUGGAAUACU---GGG------AAUGAUACUAGAGAGUCAAAUGACACAAAUAAU---------------ACUGGAAAUAUCACACUCCCAUG
B-OYI             AUAGUACUUGGAAU-----------------GAUACUACAAGGGCAAAUAGCACUGAAGUA----------------------ACUAUCACACUCCCAUG
B-SF2             AUAAUACAUGGAGG-----------------UUAAAUCACACUGAAGGAACUAAAGGAAAU---------------GAC---ACAAUCAUACUCCCAUG
B-WEAU            AUAGUACUUGGCAU-----------------GCUAAUGGUACUUGGAAGAAUACUGAAGGG------------GCAGAUAACAAUAUCACACUCCCAUG
B-YU2             -----ACUUGGAAU-----------------GAUACUAGAAAGUUAAAUAACACUGGAAGA----------------------AAUAUCACACUCCCAUG
B-pNL43           AUAGUACUUGGUUUAAUAGUACU------UGGAGUACUGAAGGGUCAAAUAACACUGAAGGA---------------AGUGACACAAUCACACUCCCAUG
D-ELI             AUAGUACAUGGAAUAUUAGUGCAUGGAAUAAUAUUACAGAGUCAAAUAAUAGCACAAACACA----------------------AACAUCACACUCCAAUG
D-NDK             AUAGUACAUGGAAUCAGACU---------AAUAGUACAGGGUUCAAUAAUGGCACA-------------------------------GUCACACUCCCAUG
O-ANT70           AUUAUACCUUUUCA------UGUAACGGAACCACCUGUAGUGUUAGUAAUGUUAGUCAA-----------------GGUAACAAUGGCACUCUACCUUG
O-MVP5180         ACUAUACUUUUAUCAAC---UGUACAAAGUCCGGAUGCCAGGAGAUCAAAGGGAGCAAUGAG------------ACCAAUAAAAAUGGUACUAUACCUUG
SIVCPZGAB         ------------------------------------------ACUGACAACAUUACAAAUGGC----------------------AUUAUAAUACUGCCAUG
```

60% less gaps



Flow diagram of `alidot`

Sorting Base Pairs By "Credibility"

1. The more sequences are non-compatible with $(i.j)$, the less credible is the base pair.

2. If the number of non-compatible sequences is the same, then the pairs are ranked by the product $\bar{p}_{i.j} \times c_{i.j}$ of the mean probability and the number of different pairing combinations.

Then we go through the sorted list and remove all base pairs that conflict with a higher ranked pair by violating conditions (i) or (ii).

Filtering Steps

1. Remove all pairs with more than two non-compatible sequences, as well as pairs with two non-compatible sequences adjacent to a pair that also has non-compatible sequences.

2. Omit all isolated base pairs.

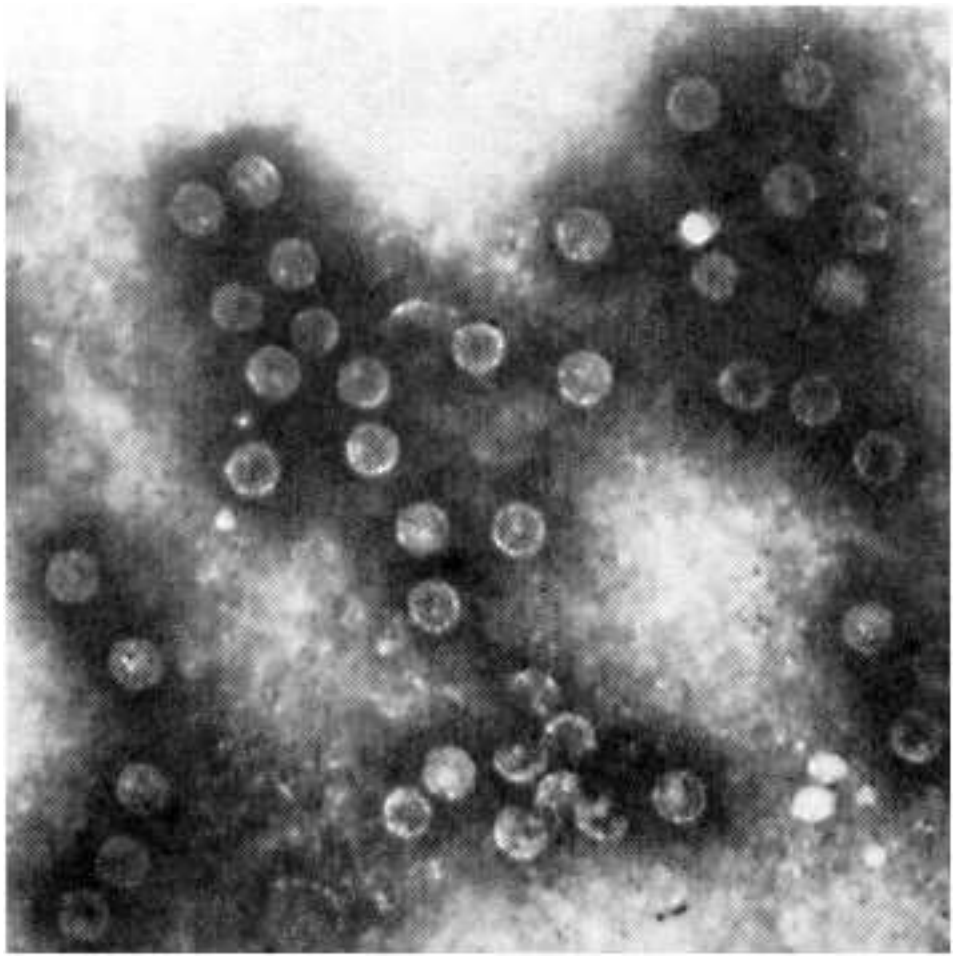
3. Collect the remaining pairs into helices and retain only those that satisfy the following conditions:

   - The highest ranking base pair must not have non-compatible sequences.

   - For the highest ranking base pair the product $\bar{p}_{i.j} \times c_{i.j}$ must be greater than 0.3.
     $\bar{p}_{i.j}$ . . . average pairing probability
     $\bar{c}_{i.j}$ . . . number of different pairing combinations.

   - If the helix has length 2, it must not have more non-compatible sequences than consistent mutations.

The remaining list of base pairs is the conserved structure predicted by the `pfrali` program.



Electron micrograph of hepatitis B virus particles, including virions, 20-nm spheres, and filaments.

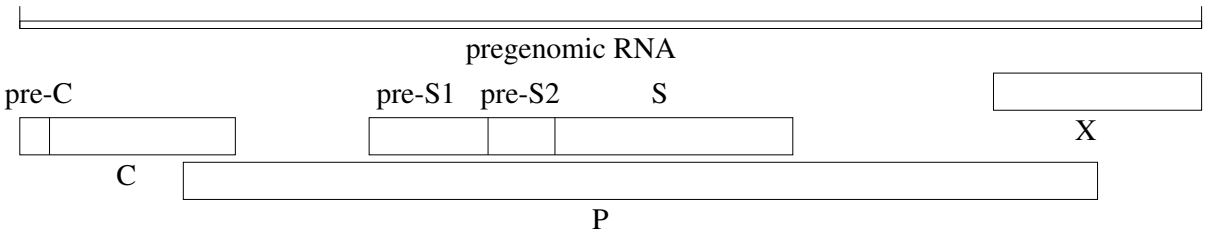


Diagram of the genome organization of hepatitis B virus with the four open reading frames (C, P, S, X).

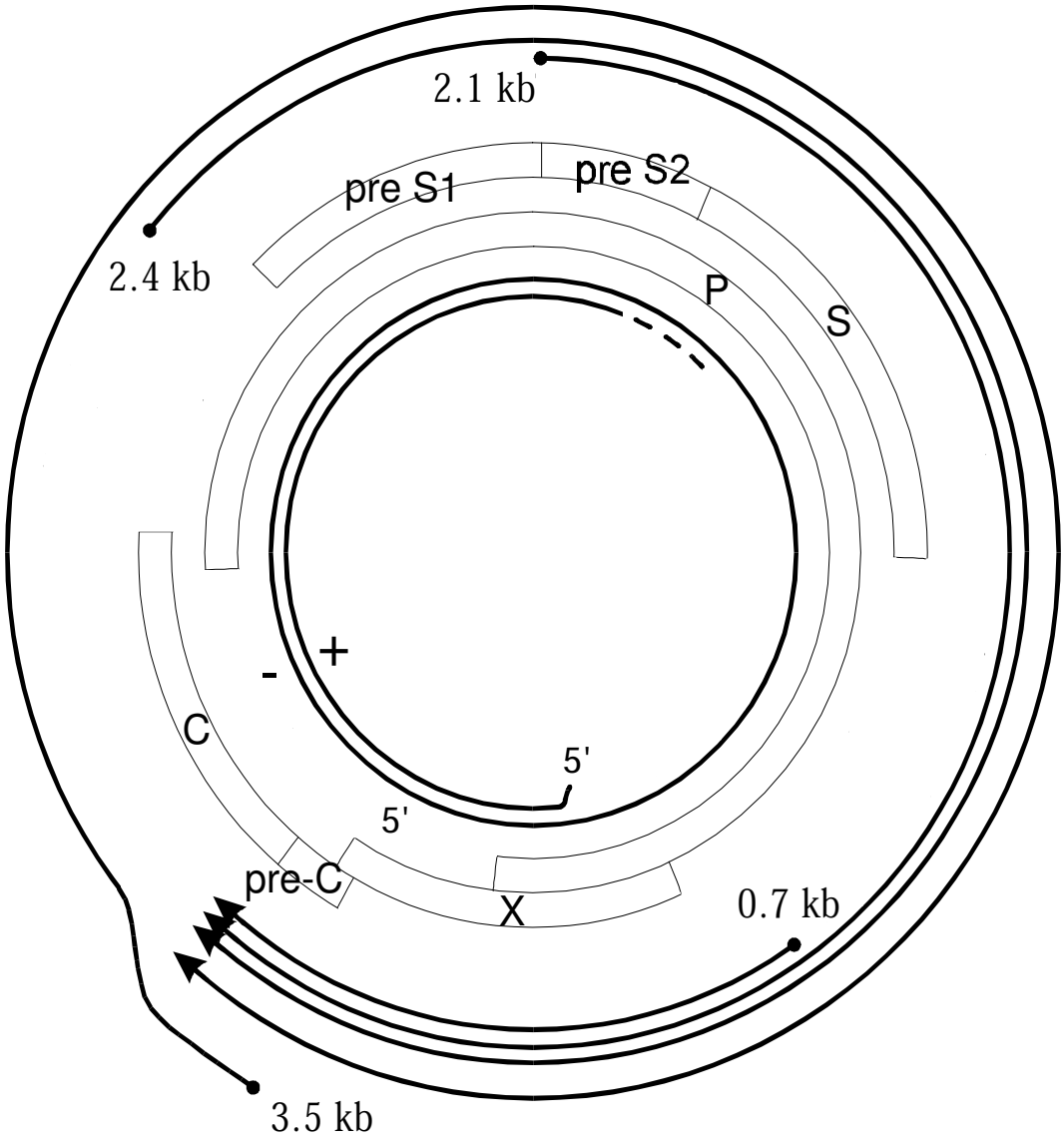
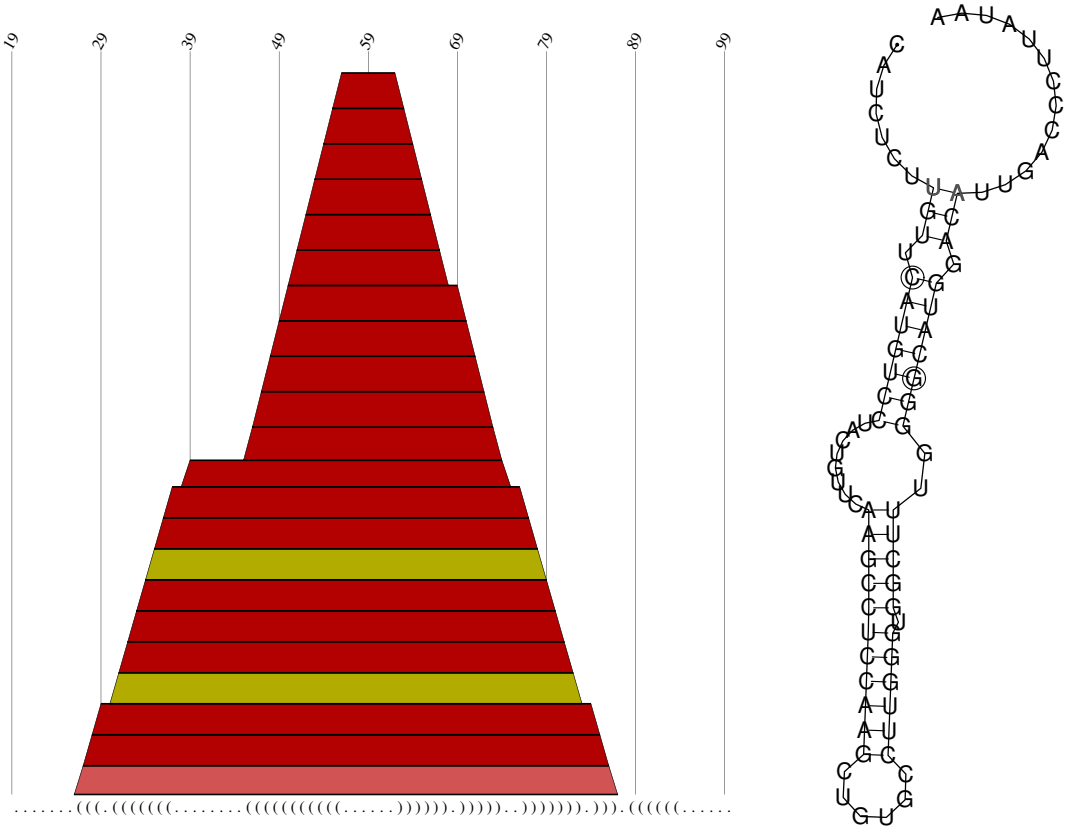


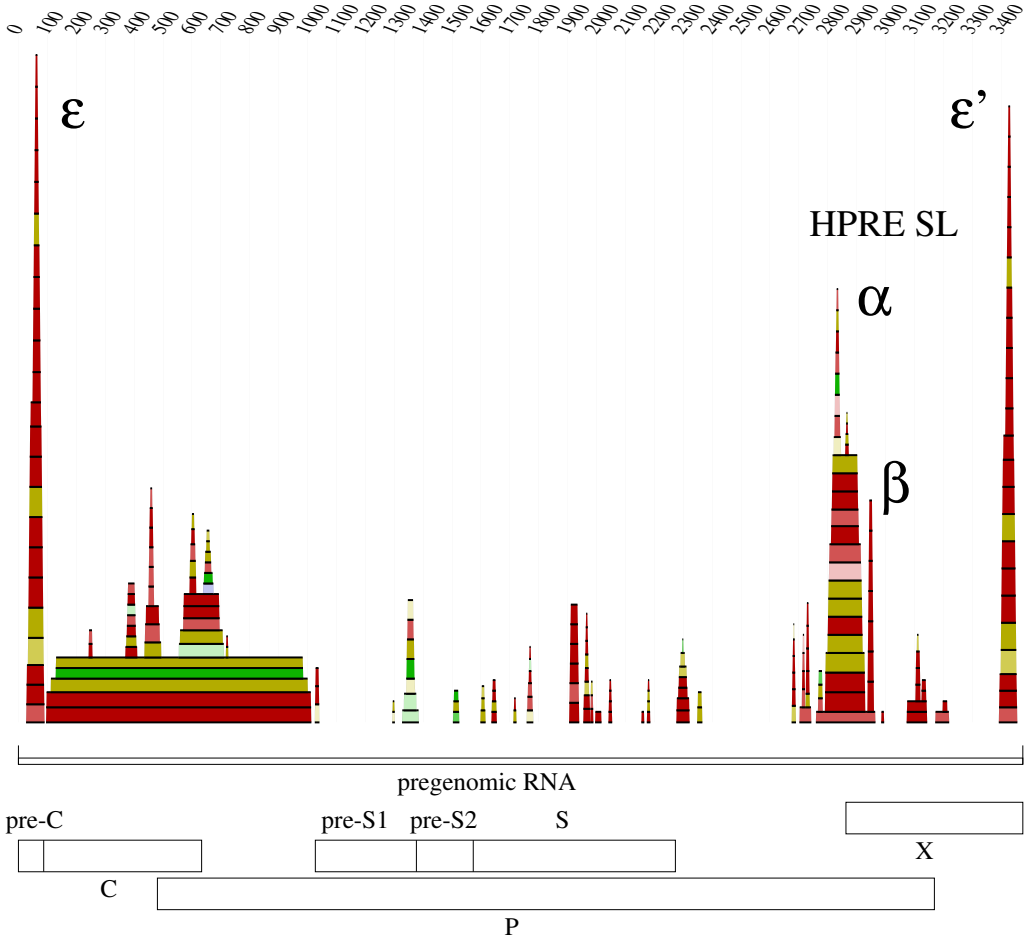
Diagram of the genome organization of hepatitis B virus indicating the DNA arrangements, the positions of the four open reading frames (C, P, S, X) and the mRNA transcripts.

Hepatitis B Virus RNA Pregenome

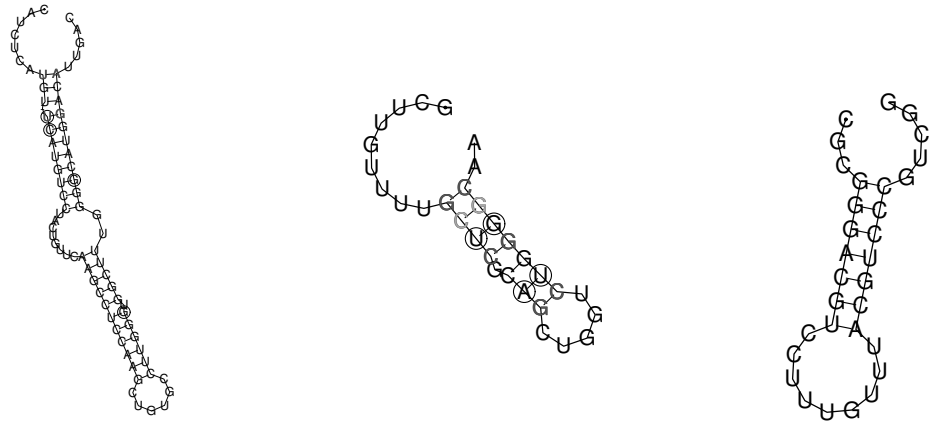


The ε-element in Human Hepatitis B Virus RNA Pregenomes.

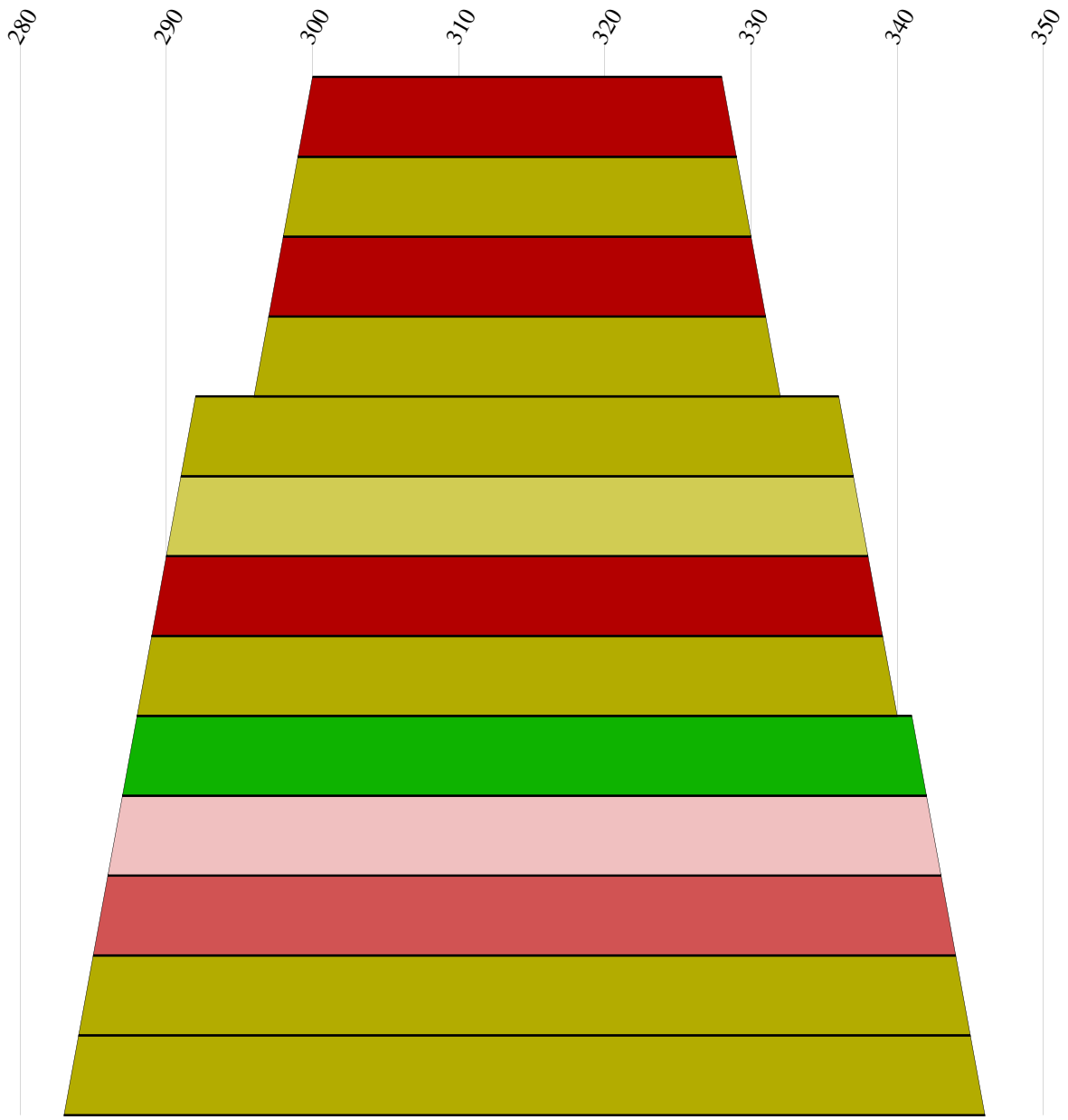
This structure is conserved among all Mammalian Hepatitis B Viruses.



The four most prominent conserved elements are the ϵ-element and its copy at the 5' and 3' end of the pregenome and the two stem loop structures α and $\beta 1$ in the HPRE region. Colors indicate compensatory mutations from 0 to 3 different types of base pairs.

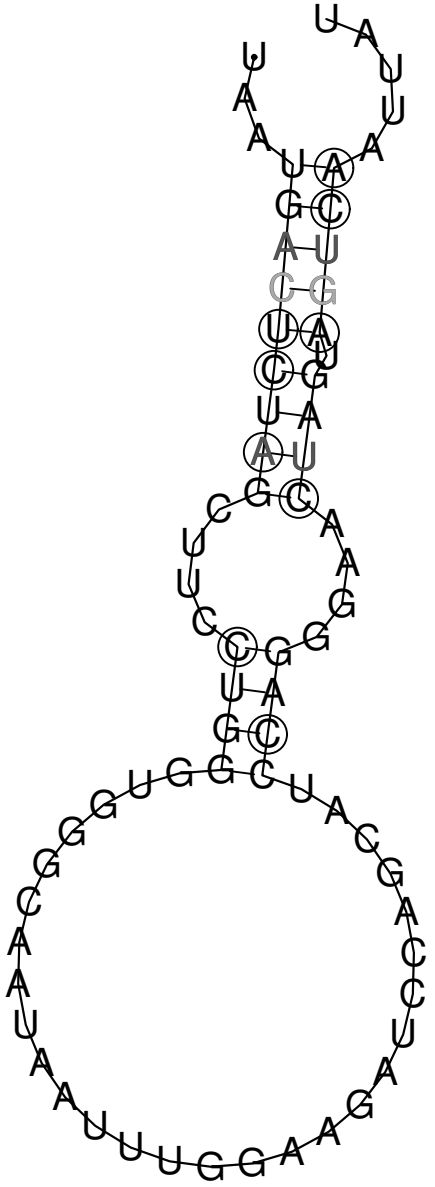


Predicted secondary structure of ϵ, α, and $\beta 1$ element. Circles indicate compensatory mutations.



...((((((((((....((((.....))))....))))..)))))).....

A secondary structure element with unknown function that appears to be conserved in the C - mRNA but does not appear as conserved structure in complete pregenomic RNA.



A secondary structure element with unknown function that appears to be conserved in the C - mRNA but does not appear as conserved structure in complete pregenomic RNA.